Ulf-Michael Stift / Thomas Schmidt

MÜNDLICHE KORPORA AM IDS: VOM DEUTSCHEN SPRACHARCHIV ZUR DATENBANK FÜR GESPROCHENES DEUTSCH

Einleitung

Gesprochene Sprache auf Tonträgern zu sammeln und der wissenschaftlichen Forschung nach entsprechender Aufarbeitung in Form von Tonaufnahmen, Metadaten und Transkripten zur Verfügung zu stellen, das sieht das „Archiv für Gesprochenes Deutsch" (AGD) im Institut für Deutsche Sprache (IDS) noch heute als seine wesentliche Aufgabe an. Diese Tradition reicht aber weit über die Gründung des IDS 1964 hinaus zurück bis in die 30er Jahre des 20. Jahrhunderts. Hier diente die Sprachsammlung zunächst den Zwecken der Dialektologie und der Phonetik, erst sehr viel später geriet auch die Gesprächsanalyse in den Blickpunkt. In diesem Beitrag zeichnen wir die Entwicklung des Archivs von seinen Anfängen in den 1930er Jahren bis zu seiner Gegenwart nach.

Eberhard Zwirner und die Gründung des Deutschen Spracharchivs (1932-1945)

Im Jahr 1932 begründete der Nervenarzt und Phonetiker Eberhard Zwirner in Berlin das „Deutsche Spracharchiv" (DSAv) als ältesten Vorläufer des heutigen AGD. Zwirner forschte seit 1928 im Institut für Hirnforschung der Kaiser-Wilhelm-Gesellschaft und initiierte dort den Aufbau einer interdisziplinären Phonetischen bzw. später Phonometrischen Abteilung. Zwirner orientierte sich an bereits bestehenden Einrichtungen im deutschsprachigen Raum wie den Phonogrammarchiven in Wien und Zürich oder der Lautabteilung der Preußischen Staatsbibliothek in Berlin. Im Gegensatz zu diesen war das Ziel des DSAv nicht die reine Archivierung und Dokumentation gesprochener Sprache und ihrer Varietäten im Bestand, sondern die Analyse konstitutiver Faktoren und der Struktur gesprochener Sprache (Knetschke/Sperlbaum 1983).

Zugute kam Zwirner die technische Entwicklung: Mit Erfindung des Kondensatormikrofons und des Niederfrequenzverstärkers wurde es möglich, über Mikrofon aufgenommene Dialoge auf Schallplatten zu speichern. So entstanden ab 1932/33 nach Testversuchen in einem Postamt in Berlin Aufnahmen in Dörfern in Brandenburg und Schlesien sowie in einem Bergwerksrevier bei Halle/Saale mit Vertretern aus allen sozialen Schichten der Bergleute (Zwirner 1983). Gleichzeitig entwickelte Zwirner in enger Zusammenarbeit mit dem (mit ihm nicht verwandten) Mathematiker Kurt Zwirner die entscheidenden Methoden der Phonometrie zur Auswertung der Sprach-Bestände.

Ab 1935 wurde unter nationalsozialistischer Herrschaft der Forschungsschwerpunkt des Instituts für Hirnforschung mit neuer Leitung anders ausgerichtet, Zwirners Abteilung 1938 aufgelöst, er selbst schied aus dem Institut aus. Das DSAv betrieb er noch ein Jahr aus privaten Mitteln in Berlin. 1939 wurde es auf Initiative des Landes Braunschweig nach Braunschweig verlegt, um dort in den neu entstehenden Industrie-Komplexen um Salzgitter und Wolfsburg mit ihren großen Populationsschmelzen synchrone Spracherhebungen durchzuführen (Knetschke/Sperlbaum 1983).
Die meisten der seit 1932 aufgenommenen Schallplatten, der zur Auswertung transkribierten Texte, Kurven und Messwerte wurden 1944 bei einem Bombenangriff auf Braunschweig vernichtet (Zwirner 1983). Einige Unterlagen zur phonometrischen Auswertung der Sprachaufnahmen, Metall-Matrizen zur Pressung von Schallplatten und Gelatine-Folien sind verstreut im AGD noch vorhanden.

Neubeginn nach dem Zweiten Weltkrieg: Das Zwirner-Korpus und verwandte Datensammlungen (1945-1971)

Nach Kriegsende 1945 begründete Eberhard Zwirner mit Unterstützung der Stadt Braunschweig und des Landes Niedersachsen das Spracharchiv und das phonometrische Forschungsinstitut („Institut für Phonometrie") neu. Die Arbeit wurde nun sehr viel stärker linguistisch ausgelegt. Neben der Unterstützung durch einen Förderverein betrieb Zwirner zur Finanzierung der Forschungsarbeit eine neurologische Praxis in Braunschweig und verfasste Gutachten (Knetschke/Sperlbaum 1983). In diesem Zusammenhang machte Zwirner auch wieder erste Sprachaufzeichnungen als Arzt-Patienten-Gespräche.
Mit der Erfindung des Tonbandes als neuem Speichermedium entwickelte Zwirner ab 1948 die Idee einer Gesamterhebung des Bestands an deutschen Mundarten und Umgangssprache auch im Hinblick auf die Veränderungen seit Ende des Zweiten Weltkriegs durch Flucht, Vertreibung und Umsiedlungen von Deutschen aus den ehemaligen deutschen Ostgebieten sowie aus Ostmittel- und Südosteuropa (Knetschke/Sperlbaum 1983). 1955 startete das DSAv in Zusammenarbeit mit zahlreichen regionalen Wörterbüchern und Universitäten die Tonbandaufnahmen für das später so benannte „Zwirner-Korpus" in der (alten) Bundesrepublik Deutschland sowie in Vorarlberg, Liechtenstein und im Elsass. Mit einigen Nacherhebungen u.a. im Kreis Herford kamen bis Anfang der 1970er Jahre fast 6.000 Aufnahmen zusammen. Der deutsche Sprachraum wurde mit einem Planquadratnetz von ca. 16 km Seitenlänge überzogen, innerhalb dieser Planquadrate wurde ein repräsentativer Ort herausgesucht und dort jeweils drei Sprecher aus verschiedenen Generationen aufgezeichnet (Zwirner 1983).

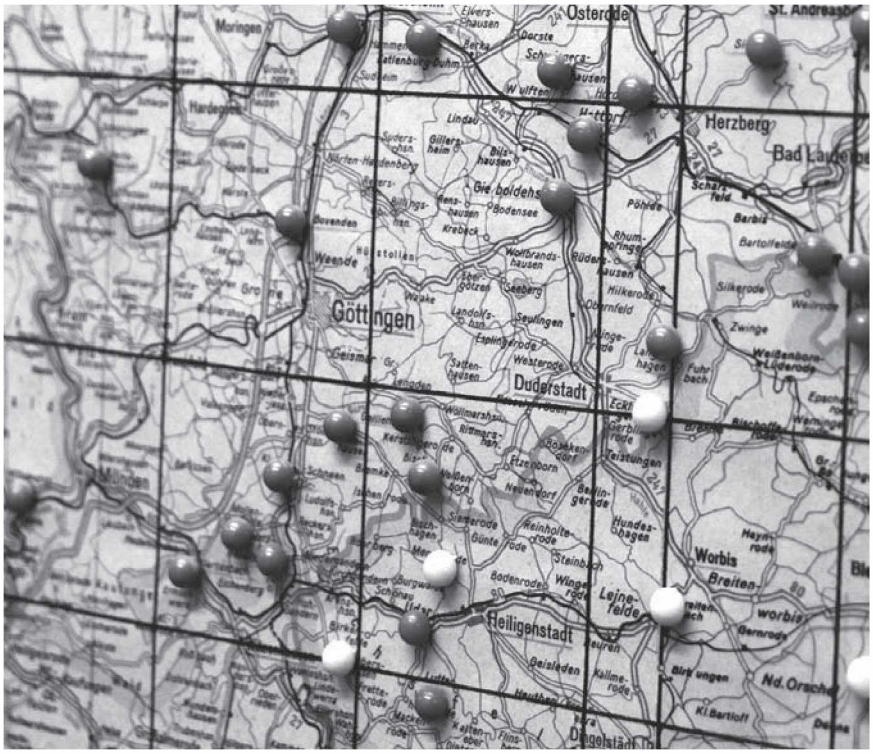

Abb. 1: Original-Landkarte mit Planquadrateinteilung und Ortspunkten aus den Korpora „Deutsche Mundarten (Zwirner-Korpus: dunkel)" und „Deutsche Mundarten: DDR (hell)"

Im Gebiet der DDR führte ab 1961 das Institut für deutsche Sprache und Literatur der Akademie der Wissenschaften in Ost-Berlin eine an Methodik, Technik und Prämissen der DSAv-Aufnahme-Aktion ausgerichtete Erhebung der Mundarten durch. Dieser Bestand von ca. 1.600 Aufnahmen wurde nach der Auflösung des damaligen Zentralinstituts für Sprachwissenschaft an der Akademie der Wissenschaften der DDR 1992 ins IDS bzw. DSAv übernommen (Wagener 2002).

Eberhard Zwirner wurde 1957 auf einen Lehrstuhl für Phonetik in Münster berufen und wechselte mit einem Teil des DSAv dorthin, in Braunschweig blieb eine Außenstelle. 1963 wechselte Zwirner bis zu seiner Emeritierung an das Institut für Phonetik der Universität Köln, das DSAv blieb in Münster und Braunschweig. Das DSAv war in den 1960er Jahren an weiteren größeren Erhebungen beteiligt: für die „Basic German"-Erhebung von deutscher Umgangssprache von J. Allen Pfeffer und W. F. Lohnes mit ca. 400 Aufnahmen in der Bundesrepublik Deutschland, in der DDR, in Österreich und in der Schweiz („Pfeffer-Korpus") übernahm es die alte Bundesrepublik, das Gebiet der DDR betreute wieder die Akademie der Wissenschaften in Ost-Berlin. In Zusammenarbeit mit dem Deutschen Sprachatlas in Marburg

entstand zwischen 1962 und 1965 eine systematische Erhebung ostdeutscher Mundarten mit fast 1000 Aufnahmen (Zwirner 1983). Diese Mundarten waren im Zwirner-Korpus an den jeweiligen Aufnahmeorten nur nach zufälliger Streuung mit erhoben worden. Die Auswertung und Verschriftung der Aufnahmen aller Korpora gelang nur zu einem geringen Teil, allein das Pfeffer-Korpus wurde komplett transkribiert. Die dialektalen Aufnahmen aus Zwirner- und Ostgebiete-Korpus sollten hochdeutsch, literarisch und phonetisch transkribiert werden, bei vielen Aufnahmen blieb es bei der hochdeutschen Umschrift (Fiehler/Schröder/Wagener 2007).

0001	**S1**	Also, erzählen Sie doch mal bitte, wo Sie geboren sind und was Ihre Eltern sind.
0002	**S2**	Ich bin am achtundzwanzigsten elften geboren in einer St/ einer Stadt bei Stettin in Pommern. Wir wurden damals nach dort evakuiert, und deshalb bin ich nicht wie meine Eltern in Stettin geboren. Mein Vater war zu der Zeit, als ich geboren wurde, nicht hier, war an der Front, und war ich mit meiner Mutter al/ ziemlich allein, ganz alleine, also wir waren .. auf/ ja auf fre/ fremde Hilfe konnten wir auch an und für sich nicht hoffen. Und/
0003	**S1**	Haben Sie keine Geschwister?
0004	**S2**	Ja, meine Geschwister, die wurden aber erst später .. geboren, in Lübeck. Meine Schwester, die ist im Jahre neunzehnhundertsiebenundvierzig geboren, und dann hab ich jetzt vorkurzem noch einen kleinen Bruder bekommen, der ist jetzt erst zwei Jahre alt.
0005	**S1**	Und was ist denn der Beruf Ihres Vaters?
0006	**S2**	Mein Vater ist .. Maschinenschlosser bei 'n/ beim TLB in Lübeck, bei der Luftwaffe, Blankensee.
0007	**S1**	Und war er das auch in Pommern, ja?
0008	**S2**	Das war er, den Beruf übte er auch in Pommern aus, ja.

Abb. 2: Transkriptausschnitt aus dem Pfeffer-Korpus (PF–_E_00011: Aufnahme mit einer in Pommern geborenen Sprecherin aus Lübeck)

Weitere Aufnahmen zur Ergänzung des Zwirner-Korpus wurden im Schwarzwald sowie in Südwestdeutschland und Vorarlberg von der 1959 begründeten Außenstelle in Tübingen unter Arno Ruoff vorgenommen, die sich 1969 als „Tübinger Arbeitsstelle Sprache in Südwestdeutschland" verselbständigte (Bethge 1976). Dazu kamen noch einige kleinere Bestände mit

hochsprachlichen, zweisprachigen und fremdsprachigen Aufnahmen. Das DSAv übernahm auch Aufnahmen auswärtiger Wissenschaftler, vor allem deutsche Mundarten in Nord- und Südamerika, Australien und europäischen Sprachinseln.

Das DSAv am IDS (ab 1971)

Nach dem altersbedingten Ausscheiden von Eberhard Zwirner infolge seiner Emeritierung sollte das Spracharchiv finanziell besser abgesichert und institutionalisiert werden. 1971 wurde es auf Empfehlung des Wissenschaftsrates unter Initiative von Hugo Moser als Präsident des Kuratoriums des Instituts für Deutsche Sprache in das IDS übernommen (Zwirner 1983), obwohl Dialektologie im engeren Sinn nicht zu dessen Arbeitsbereich zählte. Die Arbeitsstellen des DSAv in Braunschweig und Münster wurden mit anderen Einrichtungen des IDS in Bonn zusammengeführt. Die Leitung übernahm zunächst Klaus Kohler, nach dessen Wechsel nach Kiel 1972 Gerold Ungeheuer vom Institut für Kommunikationsforschung und Phonetik der Universität Bonn. Nach dessen Tod übernahmen 1976 Zwirners langjährige Mitarbeiterinnen Edeltraut Knetschke und Margret Sperlbaum die Archivleitung (bis 1990), 1979 wurde das Spracharchiv auch räumlich an den Dienstsitz des IDS in Mannheim verlegt (Knetschke/Sperlbaum 1983).

An das DSAv wurden – neben der Erhebung von zwei neuen kleineren Korpora – zunehmend Aufnahmen, Materialien und Transkripte aus abgeschlossenen Projekten des IDS abgegeben, das Sammelgebiet erweiterte sich von überwiegend dialektalen Aufnahmen über das gesamte Gebiet der gesprochenen Sprache. In dieser Zeit wurde begonnen, erstmals bis auf wenige Ausnahmen alle nach dem DSAv-Standard dokumentierten und aufbereiteten (Varietäten-)Bestände zu erfassen und zu katalogisieren (Haas/Wagener (Hg.) 1992). Fertiggestellt und veröffentlicht wurde dieser Katalog 1992 erst nach dem altersbedingten Ausscheiden von Knetschke und Sperlbaum unter der neuen DSAv-Leitung von Peter Wagener.

Gesprächskorpora: Freiburger Korpus und Korpus „Dialogstrukturen" – Die Forschungsstelle Freiburg (1965-1979)

Eine andere Tradition des AGD wurde Mitte der 1960er Jahre im IDS begründet im Zusammenhang mit der allgemeinen Tendenz in der Germanistik, sich verstärkt der grammatischen und lexikalischen Beschreibung der gesprochenen Sprache zuzuwenden (Fiehler/Schröder/Wagener 2007): Aufbau und Sammlung von Korpora unter gesprächsanalytischen und soziolinguistischen Gesichtspunkten auf Grundlage der Analyse mündlicher Kommunikation

und gesprochener Sprache im IDS. Im Rahmen des 1965 von Hugo Steger, Werner Winter, dem Institut für Deutsche Sprache und dem Goethe-Institut initiierten Projekts „Grundstrukturen der deutschen Sprache" sollten zwei Korpora aufgebaut werden, eines davon mit Texten geschriebener Sprache im IDS in Mannheim. Den Aufbau eines Korpus der alltäglichen, übergruppal und überregional verstandenen und akzeptierten gesprochenen deutschen Standardsprache (Schröder 1975) übernahm die 1966 in Kiel eingerichtete Forschungsstelle des IDS unter Leitung von Hugo Steger, die 1968 nach Freiburg[1] verlegt wurde.

Die vorhandenen linguistischen Korpus-Sammlungen des DSAv wie Zwirner-Korpus oder Pfeffer-Korpus waren für diesen Zweck nicht zu gebrauchen, weil sie eigens für sprachwissenschaftliche Zwecke zusammengestellt worden waren (Schröder 1975). Außerdem repräsentierten sie mit dem „initiierten Erzählmonolog" als spezieller Form des Interviews nur einen Typ kommunikativer Situation. Um weitere Typen der Kommunikation wie Erzählung, Vortrag, Gespräch, Diskussion, Reportage oder Unterhaltung einbeziehen zu können und spontane, alltägliche Gesprächs-Situationen zu erfassen, konnte man zum einen auf Mitschnitte von Rundfunk- und Fernsehsendungen zurückgreifen. Zum anderen war die Forschungsstelle gezwungen, die verschiedensten Gesprächssituationen im öffentlichen Bereich, vor allem aber im halböffentlichen und im privaten Bereich, über eigene Aufnahmen zu erfassen (Schröder 1975; Müller 1975). So wurden aus einem von 1967 bis 1973 zusammengetragenen Bestand von insgesamt 806 Aufnahmen, ca. die Hälfte davon Rundfunk-Mitschnitte, 222 Aufnahmen als „Freiburger Korpus" bis 1974 transkribiert.

[1] Zur Forschungsstelle Freiburg siehe auch den Beitrag von Schwitalla/Berens in diesem Band.

0001	**S2**	ja jetzt habe ich ihnen was falsches gesagt 5f5 (-) mit dem krankenhaus doktor n:n (-) da besteht kein vertrag ja 5s5 (-).
0002	**S1**	| ach so mhm.
0003	**S2**	des is| also dann also ne andere sachlage 5f5 (--) dann müßten sie also die verordnung 5g5 (--) dort abgeben ja 5s5 (-).
0004	**S1**	mhm.
0005	**S2**	9l9 und dann wird sich also (-) die klinik (-) die wird wahrscheinlich mit ihnen entweder direkt abrechnen 5s5(-) und sie reichen uns hinterher 5s5 (-) also 5g5 (--) 9l9 s is eigentlich auch nicht richtig sie müßten von anfang an (-) antrag bei uns stellen (-) auf die übernahme der krankenhauspflege 5f5 (-).
0006	**S1**	dann muß ich also zuerst einen antrag| bei ihnen stellen (-).
0007	**S2**	ja (-) ja|.
0008	**S1**	klar 5f5 (--) und hinterher (-) kann ich denn die rechnung die die (-) klinik doktor nn mir stellt (-) bei ihnen einreichen ja| zur.
0009	**S2**	die können| sie uns dann einreichen (-)| ja.
0010	**S1**	rückerstattung 5f5 (-)|.
0011	**S2**	ja.

Abb. 3: Transkriptausschnitt aus dem Korpus „Dialogstrukturen" (DS–_E_00028: Beratung bei einer Krankenkasse)

In Fortführung von Fragestellungen aus dem Grundstrukturen-Projekt betrieb die Forschungsstelle Freiburg ab 1974 ein neues Projekt zur Analyse der Organisation von Dialogen. Mit dessen Hilfe konnten Beschränkungen des Grundstrukturen-Projekts mit seinem eher statischen Begriff der „gesprochenen Standardsprache" überwunden werden, erstmals wurden auch Aspekte der Dynamik mündlicher Kommunikation untersucht (Fiehler/Schröder/Wagener 2007). Für das Dialogstrukturen-Projekt wurden 72 Aufnahmen transkribiert, die zur Hälfte dem Freiburger Bestand entnommen wurden, zur anderen Hälfte erstmals einem 1974/75 neu aufgenommenen Bestand von Video-Aufnahmen (nur die Tonspuren) vornehmlich aus dem halböffentlichen oder privaten Bereich, vor allem Beratungs- und Prüfungsgespräche. Ab Mitte der 1970er Jahre wurde die Forschungsstelle in Freiburg nach und nach aufgelöst und nach Mannheim verlegt, dennoch bildete sie die Basis für die spätere gesprächsanalytische und soziolinguistische Arbeit der Gesprochenen-Sprache-Forschung am IDS (Fiehler/Schröder/Wagener 2007).

Weitere Gesprächskorpora im IDS (1979-1989)

Nach diversen Zwischenlösungen wurde diese Arbeit ab 1979 fortgeführt in der neuen IDS-Abteilung „Sprache und Gesellschaft" (der auch das DSAv unterstellt wurde), ab 1992 geteilt in die Abteilungen „Gesprochene Sprache" (mit DSAv) und „Verbale Interaktion", 1997 wieder zusammengefasst in der heutigen Abteilung „Pragmatik". Zwischen 1979 und 1990 wurden drei große Aufnahme-Sammlungen unter gesprächsanalytischen oder soziolinguistischen Gesichtspunkten im Zusammenhang mit Projekten der Abteilung aufgebaut: Die Korpora „Beratungsgespräche", „Stadtsprache Mannheim" (mit vier Teilkorpora aus den Mannheimer Stadtteilen Westliche Unterstadt, Sandhofen, Neckarau und Vogelstang) und „Schlichtungs- und Gerichtsverhandlungen". In den 1980er Jahren wurden Teile des Freiburger Korpus und des Dialogstrukturen-Korpus sowie Kopien aus den Korpora Beratungsgespräche, Stadtsprache Mannheim und Schlichtungs- und Gerichtsverhandlungen in das DSAv übergeben (Knetschke/Sperlbaum 1983). Während Freiburger Korpus und Dialogstrukturen-Korpus bis auf wenige Ausnahmen für die wissenschaftliche Forschung zur Verfügung stehen, können von den drei anderen Korpora aus rechtlichen Gründen und wegen des Bearbeitungsstandes nur wenige (publizierte) Transkripte und die dazugehörigen Aufnahmen genutzt werden – bei diesen Projekten stand die inhaltliche analytische Arbeit im Vordergrund, nicht Archivierbarkeit und Verwendbarkeit für externe wissenschaftliche Zwecke (Fiehler/Schröder/Wagener 2007).

Umstrukturierungen und Digitalisierung (1990-2004)

Nach einer von Personalabbau und Umstrukturierungen geprägten Übergangszeit um 1990 seit dem altersbedingten Ausscheiden von Edeltraud Knetschke und Margret Sperlbaum begann mit der Übernahme der Archiv-Leitung durch Peter Wagener 1992 ein neuer Abschnitt in der Geschichte des DSAv, das sich damit endgültig von der Zwirnerschen Tradition löste. Als eines der ersten Projekte wurde der bereits erwähnte Gesamtkatalog der DSAv-Bestände fertiggestellt und veröffentlicht (Haas/Wagener (Hg.) 1992).
Schon bald ergab sich für die neue Archivleitung aus organisatorischen und technischen Gründen die Notwendigkeit zu einer umfassenden Modernisierung des Archivs. Ein Konzept wurde entworfen, dessen Umsetzung leitete ab 1994 die Digitalisierung der umfangreichen Bestände an Tonband-Aufnahmen ein und ab 1997 die Entwicklung der über Internet zugänglichen „Datenbank Gesprochenes Deutsch" (DGD) (Wagener 2002). Diese sollte zum einen die elektronische Dokumentation der Tonaufnahmen des DSAv und der dazugehörigen Transkripte und Metadaten ermöglichen, zum anderen

durch ihre Bereitstellung über das Internet die Bestände des DSAv der wissenschaftlichen Forschung außerhalb des IDS leichter zugänglich machen.
Anfang der 1990er Jahre befanden sich in den Beständen des DSAv ca. 15.000 Tonaufnahmen. Die älteren Bestände waren auf Magnet-Tonbändern gespeichert, die nach 1980 erhobenen Korpora auch auf Kompakt-Kassetten. Die Aufnahmen stammten aus Varietäten-Korpora und Gesprächs-Korpora unterschiedlicher Herkunft, aus dem Altbestand des DSAv, aus abgeschlossenen Projekten des IDS, dazu kamen eingeworbene und von auswärtigen Institutionen und Wissenschaftlern überlassene Korpora. Diese befanden sich in den verschiedensten Bearbeitungszuständen, waren höchst unterschiedlich dokumentiert und nach diversen Verfahren zu bestimmten Zwecken transkribiert. Im Zuge der Modernisierung des DSAv sollte ein einheitliches auf moderner Ton- und Computertechnologie basierendes Instrumentarium für Erschließung und Analyse aller Aufnahme-Bestände entwickelt werden.
Die Digitalisierung der Tonaufnahmen-Bestände des DSAv begann aus technischen Gründen mit den 1992 aus der DDR übernommenen Dialekt-Aufnahmen, die wegen des in den 1960er Jahren verwendeten Bandmaterials am stärksten in ihrer Existenz gefährdet waren. Als neues digitales Speichermedium verwendete man zunächst DAT-Kassetten. Später entschied das DSAv sich für ein unkomprimiertes digitales Archiv-Format auf der Basis des weit verbreiteten Speicherverfahrens für Audio-CDs (Wagener 2002). Nach und nach wurden die großen Varietäten-Korpora (Zwirner, Pfeffer, Ostgebiete), das Freiburger und das Dialogstrukturen-Korpus sowie einige kleinere Korpora digitalisiert und auf CD-Rom gebrannt. Infolge des Voranschreitens der technischen Entwicklung wurden seit 2007 die Aufnahmen als Audiofiles im WAVE-Format auf einem Server gespeichert, alle auf CD-Rom gespeicherten Aufnahmen auf den Server überspielt. 2011 konnte die Digitalisierung der Aufnahmen (zur Bestandssicherung) weitgehend abgeschlossen werden, als letztes wird das DDR-Korpus nach technischer Bearbeitung auf den Server überspielt.

Weitere neu übernommene Dialekt- und Gesprächskorpora (ab 1994)
Um dem Anspruch als „Zentrale Dokumentations- und Forschungsstelle für gesprochenes Deutsch" (Wagener 2002) gerecht zu werden, wurden seit 1994 zahlreiche Dialekt- und Gesprächs-Korpora von auswärtigen Einrichtungen und aus abgeschlossenen Projekten des IDS übernommen (und größtenteils digitalisiert). Das DSAv wollte sozusagen als Dienstleister Korpora aus öffentlich geförderten Projekten, die von diesen nicht ausreichend dokumentiert und archiviert werden konnten, einer Nutzung außerhalb der Projekte zugänglich machen. Übernommen wurden z.B. das für den Asspracheatlas von Werner König erhobene Korpus („König-Korpus"), verteilt auf drei Kor-

pora Aufnahmen aus Projekten von Anne Betten (Universitäten Eichstätt/Salzburg) zur Sprachbewahrung und Sprachentwicklung bei deutschsprachigen Emigranten aus Deutschland und Österreich in Israel und deren in Israel geborenen Nachkommen, aus dem IDS-Projekt „Sprachliche Integration von Aussiedlern" das von Katharina Meng verantwortete Teilprojekt über Zweisprachigkeit in Aussiedlerfamilien und Spracherwerb von Kindern oder als neuesten Zugang Australien-deutsche Mundarten aus dem Nachlass von Michael Clyne. Bei allen nach 1994 neu ins DSAv bzw. AGD aufgenommenen Korpora sind der Bearbeitungsstand und die rechtliche Lage höchst unterschiedlich, so dass nicht alle bisher zugänglich gemacht werden konnten bzw. manche auch gar nicht zugänglich gemacht werden können oder dürfen.

Die Datenbank Gesprochenes Deutsch (1997-2013)

Zu Verwaltung und Erschließung der Korpora auf elektronischer Basis und besserer Zugänglichkeit für die wissenschaftliche Forschung außerhalb des Instituts über Internet begann 1997 mit finanzieller Förderung der Volkswagen-Stiftung das Projekt „Computergestützte Erfassung und Erschließung der Tonaufnahmen des Deutschen Spracharchivs zum gesprochenen Deutsch" (Wagener 2002) zur Entwicklung der „Datenbank Gesprochenes Deutsch" (DGD). Das Konzept für die DGD1 wurde inhaltlich von Peter Wagener und Reinhard Fiehler, datentechnisch von Wolfgang Schneider (Dortmund) entwickelt. Mit Hilfe der DGD sollten alle Bestände des DSAv unterschiedlicher Herkunft – neben den Tonaufnahmen auch dokumentarische Daten, Transkripte und Begleitmaterialien – in einheitliche elektronische digitale Formate überführt werden. Diese Archiv-Formate sollten bei der Nutzung über die DGD miteinander in Beziehung treten und vielseitig verwendbar sein, um die Integration künftiger Neuzugänge an Korpora zu ermöglichen.

Als weitere Anforderungen galten Volltextrecherchen in Metadaten und Transkripten sowie Darstellung der Transkripte als Fließtext und in Partiturschreibung. Schließlich sollten Aufnahmen und Transkripte durch ein technisches Verfahren (Text-Ton-Alignment) so miteinander synchronisiert werden, dass zu den Textstellen im Transkript die entsprechenden Ausschnitte der Aufnahmen angehört werden können (Fiehler/Schröder/Wagener 2007). Für das Text-Ton-Alignment wurde in Kooperation mit dem Institut für maschinelle Sprachverarbeitung der Universität Stuttgart ein halbautomatisches Verfahren entwickelt. Technisch war die DGD ein Server, auf dem die Daten in Form von verknüpften XML-Seiten abgelegt wurden (Wagener 2002), keine relationale Datenbank, in Verbindung mit einem RAID-System für (freigegebene) Tondateien. Für die Volltext-Recherche in Metadaten und Transkripten wurde die im IDS für die Korpora der geschriebenen Sprache entwickelte Software COSMAS II

integriert. Nachdem zahlreiche Tonaufnahmen digitalisiert sowie Transkripte und aus dem Gesamtkatalog übernommene Metadaten in digitale Formate gebracht worden waren, startete 2002 die öffentliche Version der DGD1. Sie enthielt alle allgemeinen Informationen und Metadaten zu den eingestellten Korpora, aber nur zehn Tonaufnahmen und alignierte Transkripte als Beispiele. Dafür war der Zugang über Internet unbeschränkt.

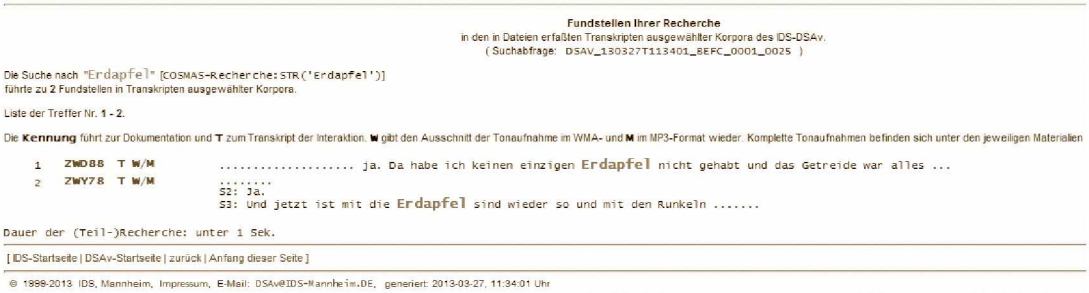

Abb. 4: Recherche nach dem Wort „Erdapfel" in der DGD

Aus datenschutz-rechtlichen Gründen wurde 2003 auf einem getrennten Server die Wissenschaftler-Version der DGD1 eingerichtet mit Zugang zu allen alignierten Transkripten und Tonaufnahmen, zur Nutzung freigegeben aber nur nach vorheriger Registrierung (Fiehler/Schröder/Wagener 2007). In der letzten Ausbaustufe (2007) waren in der Wissenschaftler-Version vollständig transkribiert und mit den Tonaufnahmen aligniert Pfeffer-Korpus, Freiburger Korpus und Dialogstrukturen-Korpus (dieses ohne Alignment). Zwirner- und Ostgebiete-Korpus waren komplett mit Metadaten, die transkribierten Aufnahmen (hochdeutsche Umschriften) mit Alignment vertreten, außerdem noch freigegebene und transkribierte Teile aus dem König-Korpus und dem ersten Korpus zum Emigrantendeutsch in Israel von Anne Betten sowie weitere kleinere Korpora mit ihren Metadaten. Unzulänglichkeiten des verwendeten Systems und der notwendige Ausbau für die Neuintegration weiterer Korpora, insbesondere von Videoaufnahmen für das gesprächsanalytische Referenzkorpus FOLK, führten ab 2006 dazu, die DGD in ein neues technisches System auf Basis einer relationalen Datenbank zu überführen. Nach Integration aller in der alten DGD1 enthaltenen Korpora in die neue DGD2 wird die DGD1 nach einer Übergangszeit mit Parallel-Betrieb endgültig abgeschaltet werden.

Das Archiv für Gesprochenes Deutsch (ab 2004)

Während einer Beurlaubung Peter Wageners für einen Forschungsaufenthalt in den USA 2000/2001 übernahm Sylvia Dickgießer die kommissarische Archivleitung. Nach Peter Wageners Rückkehr blieb er formal Archivleiter und behielt die Zuständigkeit für den Ausbau der DGD, das eigentliche Spracharchiv

wurde weiter von Sylvia Dickgießer betreut. 2004 wurde das DSAv mit dem Projektarchiv der Abteilung Pragmatik zusammengelegt, dies wurde mit der Namensänderung in „Archiv für Gesprochenes Deutsch" (AGD) unterstrichen. Damit wurden auch endgültig die originalen Bestände von Freiburger Korpus, Dialogstrukturen-Korpus und der gesprächsanalytischen Korpora Beratungsgespräche, Stadtsprache Mannheim (außer Teilkorpus Mannheim-Neckarau), Schlichtungs- und Gerichtsverhandlungen u.a. mit den Varietäten-Korpora des DSAv zusammengeführt. Nach dem Leitungs-Wechsel 2006 in der Abteilung Pragmatik von Werner Kallmeyer zu Arnulf Deppermann wurde die Arbeit in AGD und DGD mit Übernahme der AGD-Leitung durch Martin Hartung wieder zusammengeführt, Peter Wagener schied ganz aus dem AGD aus. Der Ausbau der DGD2 und der Aufbau des gesprächsanalytischen Referenzkorpus FOLK wurden forciert, zur Bestandssicherung die gesprächsanalytischen Korpora digitalisiert, auch wenn auf die zunächst geplante Verwendung für FOLK verzichtet wurde. Martin Hartung schied Ende 2010 auf eigenen Wunsch aus dem IDS aus, die AGD-Leitung übernahm kommissarisch Arnulf Deppermann. Ende 2011 kam Thomas Schmidt aus Hamburg als neuer Leiter, der bereits vorher über die Entwicklung des Transkriptionseditors FOLKER am Ausbau von DGD2 und FOLK stark beteiligt war. Als „Programmbereich Mündliche Korpora" erhielt das AGD nun eine von den anderen Projekten der Abteilung abgetrennte besondere Stellung.

Aktuelle Entwicklungen im AGD: FOLK und die DGD2

Zum 50-jährigen Jubiläum des IDS liegt der Schwerpunkt der Arbeiten im Archiv für Gesprochenes Deutsch auf zwei Projekten – dem Auf- und Ausbau des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) und der Entwicklung der Datenbank für Gesprochenes Deutsch (DGD2).

Mit der Arbeit an FOLK wird seit 2007 dem dringenden Desiderat nachgekommen, der wissenschaftlichen Öffentlichkeit ein großes Gesprächskorpus des Deutschen zur Verfügung zu stellen, das bezüglich seiner computergestützten Verarbeitbarkeit aktuellen technischen Standards genügt. FOLK strebt eine breite Stratifizierung im Hinblick auf unterschiedliche Gesprächstypen im privaten, institutionellen und öffentlichen Raum an (Deppermann/Hartung 2011). Dazu werden projekteigene Aufnahmen, ausgewählte Aufnahmen aus anderen Projekten des IDS (insb. aus dem Korpus „Deutsch Heute") sowie Datenspenden externer Projekte (z.B. aus den Projekten „Gesprochene Wissenschaftssprache kontrastiv"/Universität Leipzig und „Sprachvariation in Norddeutschland"/Universität Hamburg und weitere Universitäten im norddeutschen Raum) im Projekt nach einheitlichen Standards aufbereitet, d.h. dokumentiert, nach datenschutzrechtlichen Gesichtspunkten maskiert,

vollständig transkribiert und mit halbautomatischen computerlinguistischen Verfahren um weitere Annotationen angereichert. FOLK strebt dabei auch an, die eigene Praxis der Erhebung und Verarbeitung mündlicher Daten so zu reflektieren und zu dokumentieren, dass sie der Gesprächsforschung und verwandten Disziplinen als Beispiel einer guten Praxis dienen kann. Dazu gehört nicht zuletzt, dass – erstmals in der Geschichte des Archivs – die eingesetzten und größtenteils im Projekt selbst entwickelten Software-Werkzeuge (insb. der Transkriptionseditor FOLKER, Schmidt 2012) auch für IDS-externe Forscher und Studierende angeboten werden.

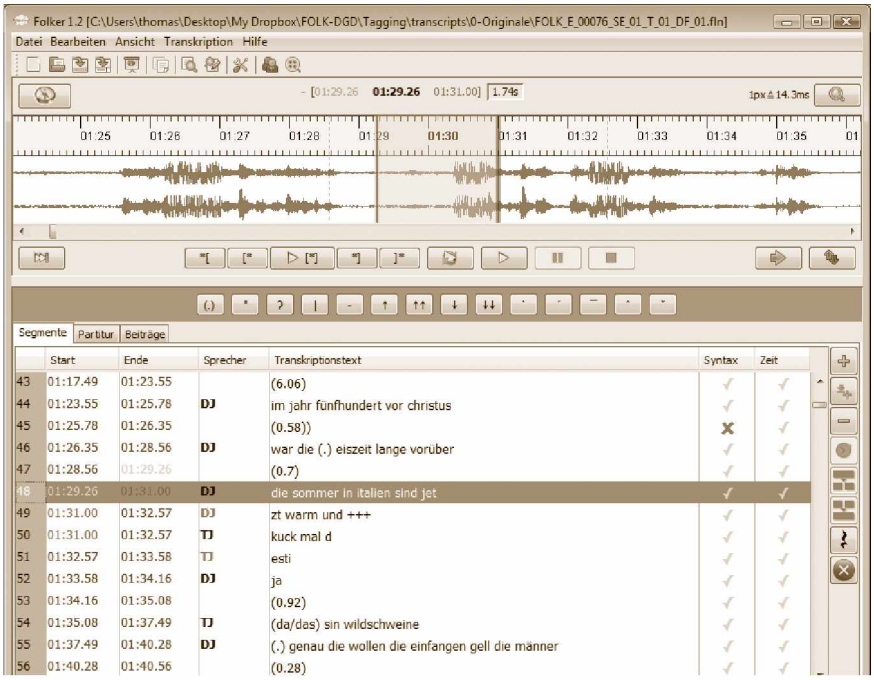

Abb. 5: Transkriptausschnitt aus FOLK im Editor FOLKER (FOLK_E_00076: Vorlesen für Kinder)

Aktuell umfasst die in der DGD2 veröffentlichte Fassung von FOLK ca. 150 Gespräche im Gesamtumfang von etwa 100 Stunden oder einer Million transkribierter Worttokens. Es ist damit bereits jetzt nach dem Zwirner-Korpus das zweitgrößte Korpus im Archiv für Gesprochenes Deutsch und soll in Zukunft jährlich um etwa 20 Stunden Aufnahmen erweitert werden.

Mit der DGD2 wird der Zugriff auf und die Arbeit mit mündlichen Korpora des Archivs auf eine neue technische Basis gestellt. Die DGD2 bietet registrierten Nutzern derzeit Zugriff auf 19 Korpora des AGD, darunter mehrere große Varietätenkorpora (Zwirner, Pfeffer etc.), die beiden „historischen" Gesprächskorpora (Freiburger Korpus und „Dialogstrukturen")

sowie FOLK. Insgesamt umfassen die über die DGD2 bereitgestellten Daten etwa 2.500 Stunden Audioaufnahmen und etwa acht Millionen transkribierte Worttokens. Wie ihre Vorgängerversion ermöglicht die DGD2 dem Benutzer zum einen ein Durchblättern der angebotenen Bestände, d.h. eine Einsicht in Metadaten zu Gesprächsereignissen und Sprechern, ein Anhören der Audioaufnahmen und ein Lesen der – größtenteils mit den Aufnahmen alignierten – Transkripte. Zum anderen erlaubt sie eine systematische Recherche auf Metadaten und Transkripten zum gezielten Auffinden einzelner Datensätze oder Belegstellen für linguistische Phänomene. Die DGD2 erfüllt damit neben der Aufgabe eines Instruments zum Zugriff auf die Archivbestände auch immer mehr die Aufgabe eines korpuslinguistischen Analyseinstruments. Durch die Umstellung auf ein modernes Datenbankmanagementsystem (Oracle) und auf eine einheitliche Handhabung von Metadaten und Transkripten auf der Grundlage von XML-Schemata ist die DGD2 nun auf eine kontinuierliche Weiterentwicklung ausgelegt. Dies betrifft sowohl die Integration neuer Daten (etwa Erweiterungen von FOLK und das Korpus „Deutsche Mundarten: DDR", s.o.) als auch die Erweiterung der Recherche-Funktionalitäten.

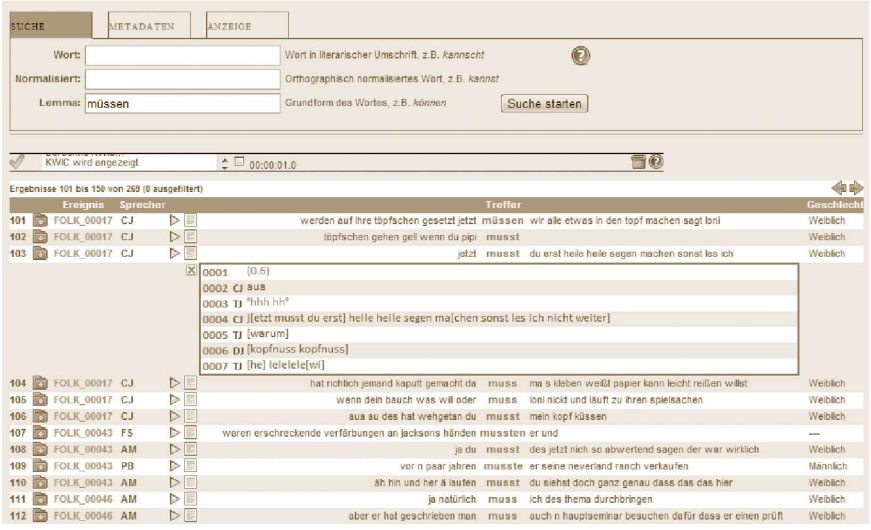

Abb. 6: Recherche nach dem Wort „müssen" in der DGD2

Ausblicke

Wie dieser kurze Abriss der Geschichte des Archivs für Gesprochenes Deutsch und seiner Vorläufer zeigt, wurde der Bereich der mündlichen Korpora zum einen durch sich wandelnde Forschungsinteressen geprägt. Waren in den frühen Tagen vor allem phonetische Fragestellungen die Motivation für die Erhebung mündlicher Daten, die sich dann auch für allgemeinere vari-

ationslinguistische Untersuchungen anboten, so bildete sich seit den späten 1960er Jahren mit der Gesprächslinguistik ein zweiter Schwerpunkt, der sich für gesprochene Sprache weniger als lautliches Phänomen, denn als Manifestation kommunikativ-interaktiver Praktiken interessiert. Das primäre Ziel von FOLK, „den „kommunikativen Haushalt" […] der deutschsprachigen mündlichen Kommunikationspraxis in seinen wesentlichen Ausprägungen [zu] repräsentieren" (Deppermann/Hartung 2011), verdeutlicht in diesem Sinne nicht nur, welche Aspekte der Mündlichkeit in den aktuellen Korpusarbeiten im Vordergrund stehen, sondern zeichnet auch eine Richtung für künftige Betätigungsfelder vor: Da die meisten Formen mündlicher Kommunikation nicht ausschließlich aus verbaler Interaktion bestehen, sondern auch kommunikativ relevante non-verbale Bestandteile enthalten (z.B. Gestik und Mimik), die sicht-, aber nicht hörbar sind, werden Gesprächskorpora, die die wesentlichen Ausprägungen der Kommunikationspraxis repräsentieren sollen, zukünftig vermehrt auf Videoaufnahmen zurückgreifen müssen. Dementsprechend wird sich auch das AGD in den nächsten Jahren auf die Verarbeitung und Bereitstellung von Videodaten einstellen.

Nicht nur in dieser Hinsicht ist die Geschichte des Archivs zum anderen untrennbar verbunden mit den technischen Entwicklungen auf dem Gebiet der Aufnahmetechnik und der Möglichkeit zur computergestützten Verarbeitung sprachlicher Daten. Die verschiedenen Phasen in der Untersuchung kindlichen Spracherwerbs („Transcripts", „Computers" und „Connectivity"), die MacWhinney (2000) respektive an der Verfügbarkeit von Tonbandgeräten, Computern und dem Internet festmacht, finden sich auch in der Geschichte des Archivs wieder. Als erster entscheidender Einschnitt kann hier der Beginn der Digitalisierung der Bestände ab den 1990er Jahren gesehen werden. Jetzt, wo die Digitalisierung der Altbestände weitestgehend abgeschlossen ist und neuere Daten (z.B. in FOLK) grundsätzlich „born digital" (also von Vorneherein mit digitalen Methoden erhoben und erschlossen) sind, ergeben sich neue Möglichkeiten und Fragestellungen durch die zunehmende Vernetzung von Sprachressourcen in digitalen Infrastrukturen. War die erste Version der DGD noch als monolithisches System konzipiert, das als alleiniger und zentralisierter Zugriffspunkt auf die Archivdaten diente, so muss die DGD2 auch ihre Anschlussfähigkeit an verteilte, service-orientierte Architekturen wie CLARIN sicherstellen. Die Daten des AGD auch auf diese Weise – z.B. durch die Bereitstellung von Metadaten in entsprechenden Online-Katalogen oder durch eine mehrere Korpus-Standorte adressierende sog. föderierte Suche – zugänglich zu machen, wird neben der Pflege und dem Ausbau der vorhandenen Komponenten eine zentrale Herausforderung in den nächsten Jahren sein.

Literatur

- **Bethge, Wolfgang** (1976): Vom Werden und Wirken des Deutschen Spracharchivs. In: Zeitschrift für Dialektologie und Linguistik (ZDL) 1/1976, S. 22-53.
- **Deppermann, Arnulf/Hartung, Martin** (2011): Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des „Forschungs- und Lehrkorpus Gesprochenes Deutsch" (FOLK) am Institut für Deutsche Sprache (Mannheim). In: Felder, Ekkehard/Müller, Marcus/Vogel, Friedemann (Hg.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen. (= Linguistik – Impulse & Tendenzen 44). Berlin/New York, S. 414-450.
- **Fiehler, Reinhard/Schröder, Peter/Wagener, Peter** (2005): Analyse und Dokumentation gesprochener Sprache am IDS. In: Kämper, Heidrun/Eichinger, Ludwig M. (Hg.): Sprach-Perspektiven. Germanistische Linguistik und das Institut für Deutsche Sprache. (= Studien zur deutschen Sprache 40). Tübingen, S. 331-365.
- **Fiehler, Reinhard/Wagener, Peter** (2005): Die Datenbank Gesprochenes Deutsch (DGD) – Sammlung, Dokumentation, Archivierung und Untersuchung gesprochener Sprache als Aufgaben der Sprachwissenschaft. Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion 6 (2005), S. 136-147 (www.gespraechsforschung-ozs.de).
- **Knetschke, Edeltraud/Sperlbaum, Margret** (1983): Das Deutsche Spracharchiv im Institut für Deutsche Sprache. 2. Aufl. (= Mitteilungen des Instituts für deutsche Sprache 6). Mannheim.
- **Haas, Walter/Wagener, Peter** (Hg.) (1992): Gesamtkatalog der Tonaufnahmen des Deutschen Spracharchivs (= Phonai 38 und 39). Tübingen.
- **MacWhinney, Brian** (2000): The CHILDES project: tools for analyzing talk. Mahwah, NJ u.a.
- **Müller, Rolf** (1975): Die Konzeption des Corpus gesprochener Texte des Deutschen in der Forschungsstelle Freiburg des Instituts für Deutsche Sprache. In: Gesprochene Sprache. Bericht der Forschungsstelle Freiburg. 2. Aufl. (= Forschungsberichte des Instituts für deutsche Sprache 7). Tübingen, S. 47-75.
- **Schmidt, Thomas** (2012): EXMARaLDA and the FOLK tools. In: Proceedings of LREC, ELRA.
- **Schröder, Peter** (1975): Die Untersuchung gesprochener Sprache im Projekt ‚Grundstrukturen der deutschen Sprache' – Planungen, Probleme, Durchführung. In: Gesprochene Sprache. Bericht der Forschungsstelle Freiburg. 2. Aufl. (= Forschungsberichte des Instituts für deutsche Sprache 7). Tübingen, S. 5-46.
- **Wagener, Peter** (2002): Gesprochenes Deutsch online. Zur Modernisierung des Deutschen Spracharchivs. In: Zeitschrift für Dialektologie und Linguistik (ZDL) 3/2002, S. 314-335.
- **Zwirner, Eberhard** (1983): Fünfzig Jahre Deutsches Spracharchiv. In: Zeitschrift für Dialektologie und Linguistik (ZDL) 1/1983, S. 35-43.